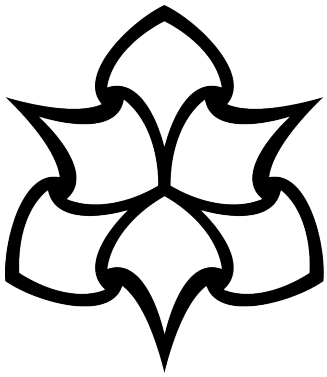


The development of a video retrieval system using a clinician-led approach

John Darby[a,*], María B. Sánchez[b], Sarah Bew[c], Ian Loram[b], Penelope Butler[b]

[a]*School of Computing, Mathematics & Digital Technology, Manchester Metropolitan University, Faculty of Science & Engineering, John Dalton Building, Chester Street, Manchester M1 5GD, United Kingdom*
[b]*Research Centre for Musculoskeletal Science & Sports Medicine, Manchester Metropolitan University, Faculty of Science & Engineering, John Dalton Building, Chester Street, Manchester M1 5GD, United Kingdom*
[c]*The Movement Centre, Oswestry, Shropshire SY10 7AG, United Kingdom*

Abstract

Patient video taken at home can provide valuable insights into the recovery progress during a programme of physical therapy, but is very time consuming for clinician review. Our work focussed on i) enabling any patient to share information about progress at home, simply by sharing video and ii) building intelligent systems to support Physical Therapists (PTs) in reviewing this video data and extracting the necessary detail. This paper reports the development of the system, appropriate for future clinical use without reliance on a technical team, and the clinician involvement in that development. We contribute an interactive content-based video retrieval system that significantly reduces the time taken for clinicians to review videos, using human head movement as an example. The system supports query-by-movement (clinicians move their own body to define search queries) and retrieves the essential fine-grained movements needed for clinical interpretation. This is done by comparing sequences of image-based pose estimates (here head rotations) through a distance metric (here Fréchet distance) and presenting a ranked list of similar movements to clinicians for review. In contrast to existing intelligent systems for retrospective review of human movement, the system supports a flexible analysis where clinicians can look for any movement that interests them. Evaluation by a group of PTs with expertise in training movement control showed that 96% of all relevant movements were identified with time savings of as much as 99.1% compared to reviewing target videos in full. The novelty of this contribution includes retrospective progress monitoring that preserves context through video, and content-based video retrieval that supports both fine-grained human actions and query-by-movement. Future research, including large clinician-led studies, will refine the technical aspects and explore the benefits in terms of patient outcomes, PT time, and financial savings over the course of a programme of therapy. It is anticipated that this clinician-led approach will mitigate the reported slow clinical uptake of technology with resulting patient benefit.

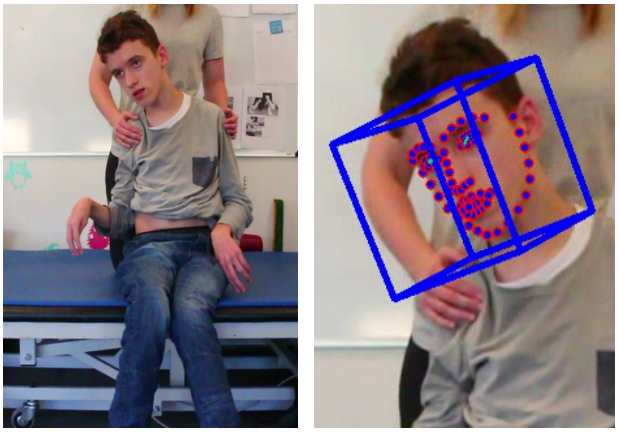
Keywords: Search methods, Telemedicine, Computer vision, Videoconferences, Video sharing.
2010 MSC: 00-01, 99-00

1. Introduction and Related Work

1.1. Introduction

Expert and intelligent systems have been shown to have considerable value to Physical Therapists (PTs) working to help improve movement and function in patients with injuries or long-term disabilities. There have been two major groups of system used: 'proactive' and 'retrospective'. Proactive systems use sensors

*Corresponding author; Tel. 011 44 161 247 1542
**Declarations of interest: none
Email addresses: j.darby@mmu.ac.uk (John Darby), m.sanchez.puccini@mmu.ac.uk (María B. Sánchez), sarahbew@the-movement-centre.co.uk (Sarah Bew), i.loram@mmu.ac.uk (Ian Loram), pennybutler2015@outlook.com (Penelope Butler)



(a) (b)

Figure 1: Extracting pose-based features from images: (a) This patient with cerebral palsy cannot achieve an upright head posture; (b) An estimate of 3D head pose (blue box) extracted directly from the image using the OpenFace (Baltrušaitis et al., 2016) computer vision library.

to extract parametric representations of body pose and respond to it in real-time. They have enabled PTs to connect with patients' progress at home, acting 'on behalf of' the PT and providing motivation through gamification (Saini et al., 2012; Alankus & Kelleher, 2015; Mortazavi et al., 2016), or feedback and guidance through visualisation (Uzor & Baillie, 2013; Tang et al., 2015). For engaged and cooperative patients, these fully-automatic expert systems have made possible impressive advances, including the ability to customise target movements (Alankus et al., 2010), explicitly addressing compensatory strategies (Alankus & Kelleher, 2015), dynamically responding to changes in patient ability (Mortazavi et al., 2016), and remotely summarising progress to clinicians (Kimel, 2005).

Retrospective systems provide an alternative to fully-automatic expert systems for progress sharing: these systems use sensors to extract parametric representations of body pose and store it for offline analysis by PTs. For example, patients wear a 'smart cap' to record head movements (Huang et al., 2014) or a 'smart sleeve' to record arm movements (Ploderer et al., 2016b). The challenge is then to develop intelligent systems which let PTs make fast and efficient searches of the resulting data logs for clinically relevant movements. These approaches show great potential as a way to monitor movements made during both exercises (Huang et al., 2014) and daily life (Ploderer et al., 2016b), while making minimal demands on the patient.

1.2. Motivation for the Current Study

Although the value of proactive and retrospective systems has been demonstrated, not all patients are readily able to cooperate. Problems can include difficulty in seeing/understanding rules of gameplay, or interpreting other forms of computerised feedback, and unhappiness about having to wear invasive sensors. An example is the young patient with complex neuromuscular control problems, such as a child with Cerebral Palsy who lacks head control (Fig. 1a). The same constraints also apply to other disabilities and age groups (e.g., adult stroke or injuries that affect cognition as well as movement).

A further difficulty is that previous studies of retrospective systems have revealed that PTs can find the stored parametric representations of body pose difficult to interpret in isolation (Ploderer et al., 2016a). For example, PTs interested in head movements reported difficulty in interpreting metrics like rotational velocity in the absence of any context (Huang et al., 2014).

Seeing the original context of movements – for example in videos – can reveal subtle but important information such as contact between limbs or the gradual shifting of weight distributions (Aggarwal et al., 2016) or complicating external factors like contact with external surfaces, or physical assistance from a carer (Ploderer et al., 2016b). Video cameras avoid the need for patients to wear invasive sensors while preserving

the context of their movements for PTs to consider. All of these factors are vital to accurate and effective physical therapy.

The primary motivation for this work is thus the need:

i) to enable any patient to share information about progress at home, simply by sharing video (with assistance from carers, if needed)

ii) to build intelligent systems to support PTs in reviewing the resulting video data and extracting the detail relating to the body movements of interest for a given patient: i.e., video-based retrospective systems.

We have taken the problem of head control as an example but worked to ensure our proposed methods have potential applications to movements in all parts of the body.

There is now a growing array of consumer devices which record encrypted video directly to secure Cloud-based storage and which have the potential to support this kind of video sharing conveniently, and at scale (e.g., Nest Cam, Netgear Arlo, Amazon Cloud Cam). As an example, the Netgear Arlo – used in this study – is small, completely wireless, has HD resolution, universal mounting threads, and a magnetic casing, and is suitable for mounting on a patient's wheelchair or standing frame.

1.3. Automated Analysis of Human Movement in Video

Two other groups of expert systems come into play when considering the automated analysis of human movements from video: systems for human action recognition from videos (Poppe, 2010; Herath et al., 2017) and systems for content-based video retrieval (Geetha & Narayanan, 2008; Hu et al., 2011). These systems and their potential bearing on the problem of progress sharing and the extraction of relevant detail are further discussed in the following sections.

1.3.1. Human Action Recognition

Human action recognition systems aim to recognise individual human actions from video footage featuring complete action executions (Poppe, 2010; Herath et al., 2017). Modern Convolutional Neural Network (CNN) approaches based on local appearance-based and motion-based features (Simonyan & Zisserman, 2014; Yue-Hei Ng et al., 2015) are able to recognise coarse actions (e.g., 'walk', 'stand up') in challenging real-world footage that may include very varied compositions, camera movements, multiple people, occlusions, etc. However, these 'low-level' feature-based approaches are not as effective in problems involving fine-grained human action recognition (e.g., Ni et al. (2014); Rohrbach et al. (2012)) where differences in actions of interest can be quite subtle, including differences in the performance of the same action. This subtle detail is a requirement of our proposed system.

Progress in recognising fine-grained actions in videos has been made by explicitly considering 'high-level' features relating to the human within the images (Kläser et al., 2010; Yao et al., 2010; Chéron et al., 2015; Choutas et al., 2018; Luvizon et al., 2018; Negin et al., 2018). An example of this is performing automatic human pose estimation in order to extract appearance-based features from specific body parts, rather than across the entire image (Chéron et al., 2015). In fact, where pose estimation can be reliably performed, results suggest that using *only* pose-based features (e.g., joint locations, distances between joints, relative velocities) can allow for equally good or even superior performance to using appearance-based features in fine-grained action recognition problems (Yao et al., 2011).

Inspired by this finding and by recent advances in the accuracy and robustness of image-based pose estimation (Guo et al., 2016), we take an approach to the analysis of fine-grained human movements which is based entirely on the results of human pose estimation; specifically, head pose estimation (Murphy-Chutorian & Trivedi, 2009; Fanelli et al., 2012). A contemporary method for image-based head pose estimation (OpenFace, Baltrušaitis et al. (2016, 2013, Fig. 1b)) is used to extract pose-based features *in the clinic*. The pose-based features (here a sequence of quaternions) can then be used as a signal for the identification of relevant actions by a retrospective clinical software system, while the original video remains available for context if needed during clinician review. Simultaneously, patients are freed from the requirement to have bespoke software and hardware in their home.

The traditional approach to action recognition is to collect examples from each of a number of different actions of interest, extract features, and train a classifier to distinguish between them (Yao et al., 2011; Simonyan & Zisserman, 2014; Chéron et al., 2015). In the context of physical therapy, the need to provide sufficient numbers of training examples in advance limits flexibility when PTs use the system to analyse new videos: they can only retrieve actions which fall into previously defined classes. In practice, different groups of PTs will be interested in different movements (e.g., based on the type of therapy they use) and may have unique questions to ask about the movements of individual patients (e.g., based on the nature of their disability, or their progress within the programme of therapy). A more flexible approach to defining actions of interest is a requirement of our proposed system.

1.3.2. Content-Based Video Retrieval

Low-level feature-based approaches to recognition have also been employed in the wider field of content-based video retrieval, where users can type a natural language query in order to retrieve videos with matching content from a database of target videos (Geetha & Narayanan, 2008; Hu et al., 2011). State-of-the-art systems (e.g., Ueki et al. (2016, 2017)) work by recognising basic 'concepts' (which can include, but are not limited to, human actions) from low-level features within the images comprising each target video: e.g., scenes, objects, people, actions, relationships. A user's query is then mapped to concept names using semantic similarity techniques, and videos are retrieved based on their aggregated scores across the relevant concept classifiers (Awad et al., 2017).

Results from the field of action recognition suggest that by considering high-level features, a clinical concept-based video retrieval system could be trained to include fine-grained human actions as concepts. However, training data must still be available for each action of interest in advance, and clinicians must either agree a standard set of action descriptions to ensure they type effective search queries, or rely on automatic semantic similarity measures (Ueki et al., 2017; Mikolov et al., 2013). Data requirements for training concept classes featuring human actions are substantial, with most methods relying on large research-community generated datasets (e.g., Soomro et al. (2012)), and the difficulty of mapping natural language queries to concepts can be seen in the reduced performance of fully-automatic systems (which process natural language queries verbatim) versus manually-assisted systems (where keywords are manually selected based on natural language queries) (Awad et al., 2018).

The challenges of providing sufficient training examples and defining effective search queries in words has seen the investigation of alternatives to text-based queries which use richer forms of query data, and direct comparisons between query data and target videos via more general purpose feature-based similarity measures which bypass concept learning altogether. Examples include: *query-by-example* where users can provide an example image or video defining their query (e.g., Snoek et al. (2007); Yang et al. (2013)), *query-by-sketch* where users can draw their query (e.g., Hu et al. (2007)), *query-by-object* where users can provide an image of an object defining their query (e.g., Sivic & Zisserman (2003)), and combinations thereof (e.g., Moreno-Schneider et al. (2017)).

Here we argue that the most natural method for defining queries pertaining to fine-grained human movement is to move your own body: *query-by-movement*. We then make a direct comparison between high-level pose-based features extracted from a user's query and high-level pose-based features extracted from target videos. This approach gives individual PTs the freedom to look for any movements they require, based on the type of therapy they use, the patient's disability and/or the patient's level of progress.

1.3.3. Interactive Content-Based Video Retrieval

Despite the potential to define rich, multimodal queries, fully-automatic content-based video retrieval remains a very challenging problem in the general case (Lokoč et al., 2018). This has led to an interest in *interactive* content-based video retrieval systems which give users some degree of control over the search processes which operate following submission of a query, and the final decision on the relevance of videos which are retrieved (Schoeffmann et al., 2010, 2015).

Interactive systems range from carefully crafted user interfaces (UIs) which support users in fast manual browsing of many hours of video via keyframes (Hürst et al., 2015), to more complex methods which can dynamically cluster or re-rank keyframes based on visual similarity or highlighted concepts (Barthel et al.,

Figure 2: A PT's query for a left-to-right head rotation, recorded in the clinic (left), produces a ranked list of similar movements from videos that have been recorded at home (right).

2015; Lu et al., 2017), or filter keyframes based on colours or objects present (Bailer et al., 2016; Moumtzidou et al., 2017). Ultimately however, each approach aims to support the user in quickly identifying promising scenes before watching them back to make judgements about their relevance. Though interactive systems typically take longer than fully-automatic systems to achieve equivalent levels of recall, they can be used to ensure high precision results (Lokoč et al., 2019).

Interactive systems are also effective when users do not have a detailed search term in mind, or are interested in more abstract properties which are true for large portions of the target video database (Lokoč et al., 2018). The interactive, human-in-the-loop approach still allows for relevant results to be identified quickly, and even though they may differ substantially from an initial query.

In our own application scenario, where PTs must consider subtle aspects of context which cannot be captured by pose-based features in isolation, we argue that it is natural to formulate a solution based on interactive search, where PTs are supported in quickly reviewing videos generated by an automated search process in order to make final judgements about clinical relevance.

1.4. Proposed System

We propose an interactive video retrieval system which is specifically aimed at allowing PTs to retrieve fine-grained human head movements, based on the output of a contemporary head pose estimation technique. One useful side benefit of performing pose estimation in the clinic is that clinicians can use it to extract pose-based features from videos of themselves. We draw on this fact to provide a query-by-movement interface where PTs can move their own head in order to initiate searches for any defined movement, thus facilitating a more flexible clinical analysis than has been possible with previous retrospective approaches (Ploderer et al., 2016b; Huang et al., 2014). We do this by implementing a novel algorithm for identifying and ranking similarities between sequences of head pose estimates and using it to build an interactive video retrieval UI specifically for PTs who train head control, using a participatory design process.

PTs can use the system to define new search queries simply by performing the desired head movement in front of a webcam at the clinic, and then use their queries to retrieve ranked lists of similar movements in videos that have been recorded at home (Fig. 2). After submitting a query PTs see keyframe summaries of each video in the ranking, allowing them to make quick and effective decisions about which videos to watch back. They can then play and re-watch videos to consider context, 'flag' relevant videos with different colour codes, and save their search results.

In common with other interactive search methods, PTs can also use the system to browse movements where their interests are less well defined. For example, rather than stopping searching as soon as movements begin to depart significantly from their original search query, they can continue their search to see what form these departures take and their relative rates of occurrence (e.g., where a patient is struggling with a particular exercise or a commonly performed movement, see the lower-ranked result in Fig. 2).

Finally, to address requirements identified by other investigations into retrospective methods (Huang et al., 2014; Ploderer et al., 2016b), we also provide PTs with quick search summaries, showing the total

number of events they have flagged for a patient, the times and dates they occurred, and their distribution over the programme of therapy.

1.5. Clinician-Led Study

Clinical uptake of technology enabled care methods, such as retrospective methods, has generally been relatively slow, with cultural resistance from clinicians who report feeling they are victims of 'technology push', excluded from system design processes, and underprepared and undertrained for new deployments (European Commission, 2014). Bringing about sustained changes in the delivery of physical therapy is non-trivial; Hochstenbach-Waelen & Seelen (2012) propose use of a 5-stage process which focuses on *clinicians first*, with successive phases of 'orientation' (creating awareness, interest and support) and 'insight' (knowledge, understanding of relationship to current practices, and potential benefits) before 'acceptance', 'change' and 'sustained change' become possible.

Our work at this stage of project development was entirely focused on clinicians and clinician-led, with no patient involvement. PTs took cameras into their own homes to generate test data, helped to design the analysis software through a participatory design process, and evaluated the performance of the final system by reviewing their colleagues' videos. This approach was taken in order to ensure the orientation and insight components were appropriate for clinical staff and to leave them in a position where they could adopt the methods for use with patients, should they wish to, without reliance on a technical team.

1.6. Contributions

The paper makes the following specific contributions:

- A search algorithm for comparing a short sequence of head pose estimates extracted from video of a clinician with longer sequences extracted from video recorded at home, in order to produce a ranked list of similar movements.
- A search system which allows clinicians to review those movements, with the original video for context, record judgements about their clinical relevance, and view summaries of their searches.
- Empirical evidence from clinician-led evaluation of the search system by a group of PTs who train head control in children with neurodisability.

In terms of the final system's novelty versus other existing expert and intelligent systems, to the best of our knowledge it represents:

- The first retrospective progress monitoring system to preserve context through video;
- The first content-based video retrieval method to support query-by-movement;
- The first content-based video retrieval method to support fine-grained human actions.

2. Methods

2.1. Overview

The proposed system allows clinicians to define search queries by recording a 'query video' of themselves making the head movement they are interested in finding in the 'target video' that has been recorded at home. We use a commercial Netgear Arlo system (Netgear, 2018) for secure and convenient recording at home (Fig. 3), with only a standard webcam at the clinic. The OpenFace (Baltrušaitis et al., 2016) face tracker (Baltrušaitis et al., 2013) is used to extract estimates of 3D head rotation from both query video and target video, and a clustering-based algorithm is used to find approximate matches between the resulting sequences of rotations. Matches are ranked based on their similarity (using discrete Fréchet distance comparisons between sequences of quaternions) and displayed back to the clinician as short videos, preserving the context of the original movement, and giving the PT the final decision on clinical relevance

Figure 3: Wireless Netgear Arlo camera used for home recording; one for front view and one for side view.

(Fig. 2). To address requirements identified by other investigations into retrospective methods (Ploderer et al., 2016b; Huang et al., 2014), the system also provides PTs with quick search summaries, showing the total number of events they have flagged as relevant, the times and dates they occurred, and their distribution over time.

2.2. Search Algorithm

This algorithm provides a method for comparing a sequence of 3D head rotations from a query video (the 'query data') with a longer sequence of 3D head rotations from a target video (the 'target data'), in order to generate a ranked list of similar movements. The approach consists of the following stages: (i) clustering of the target data; (ii) extraction of similar movements from the clustered target data given new query data from the PT; (iii) scoring and ranking of results for presentation to the PT.

2.2.1. Clustering Rotations

To cluster 3D head rotations we use a unit quaternion representation $\mathbf{q} = (q_1, q_2, q_3, q_4)^\top$ where rotations can be thought of as points on the surface of a 4D unit sphere, and the following metric for computing the distance, ψ, between two rotations, $\mathbf{q}_1$ and $\mathbf{q}_2$, is computationally efficient and accounts for the fundamental quaternion ambiguity $\mathbf{q} = -\mathbf{q}$ (Huynh, 2009),

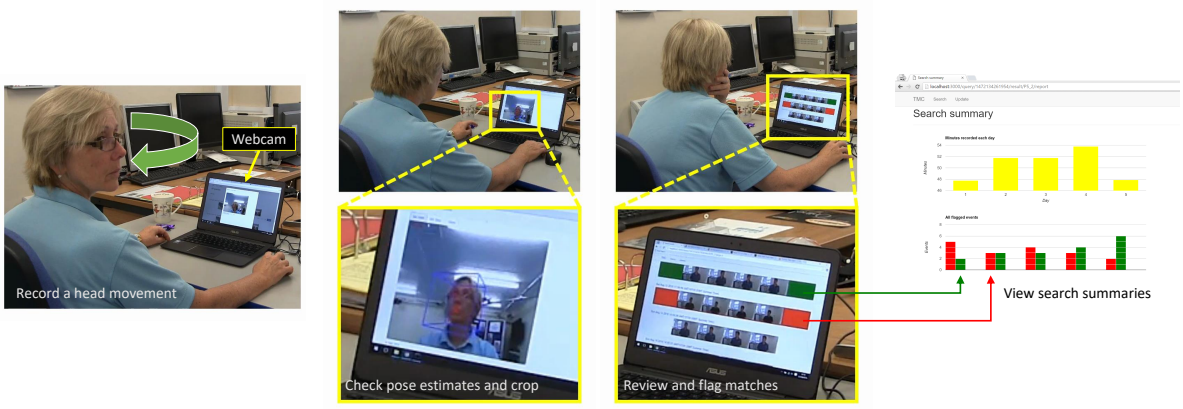
$$\psi(\mathbf{q}_1, \mathbf{q}_2) = \arccos(|\mathbf{q}_1.\mathbf{q}_2|). \tag{1}$$

We use the method of Yershova et al. (2009) to generate a set of N quaternion cluster centres, uniformly distributed over the space of all rotations, $C = \{\dot{\mathbf{q}}_1, \dot{\mathbf{q}}_2, ..., \dot{\mathbf{q}}_N\}$, and use these points to perform nearest neighbour clustering of target data in terms of the distance measure in Equation (1). Target data traces out a trajectory across the surface of the unit sphere over time, $P = \{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_T\}$, and the ith cluster captures a number of short, consecutive runs of data where the trajectory passes nearby: $\mathcal{R}_i = \{R_1, R_2, ..., R_{r_i}\}$, where $R_j = \{\mathbf{q}_{t_j}, \mathbf{q}_{t_j+1}, \mathbf{q}_{t_j+2}, ..., \mathbf{q}_{T_j}\}$. We formulate our approach to searching P in terms of the set of all runs $\{\mathcal{R}_i\}\forall i \in C$ which are mutually disjoint and cover P.

2.2.2. Generating Results

Results are generated based on query data from the PT consisting of a new sequence of head rotations, $S = \{\hat{\mathbf{q}}_1, \hat{\mathbf{q}}_2, ..., \hat{\mathbf{q}}_M\}$. The first and last quaternions $\hat{\mathbf{q}}_1$ and $\hat{\mathbf{q}}_M$ are used to find a subset of starting clusters $C_{\text{start}} \in C$ and stopping clusters $C_{\text{stop}} \in C$ that lie within a fixed distance δ of $\hat{\mathbf{q}}_1$ and $\hat{\mathbf{q}}_M$ respectively, in terms of Equation (1). Any continuous sequence of rotations in P that connects a cluster from C_{start} with a cluster from C_{stop} is then selected as a result, subject to the following two conditions: i) that the total time elapsed between leaving the start cluster and entering the end cluster falls within some acceptable margin of the original query duration (here we use 50-150%); ii) that for any pair of overlapping results, the one with the highest discrete Fréchet distance from the query data is deleted (see following section).

During a search these rules can be used to generate results quickly by iterating over the set of all runs in the starting clusters, $\mathcal{R}_{\text{start}} = \{\mathcal{R}_i\}\forall i \in C_{\text{start}}$ and comparing the timestamps of their last members against those of the first members in the equivalent set of runs $\mathcal{R}_{\text{stop}} = \{\mathcal{R}_i\}\forall i \in C_{\text{stop}}$.



(a) (b) (c) (d)

Figure 4: Overview of search system: (a) A PT moves their head to record a query video using a laptop webcam; (b) The PT reviews the accuracy of the pose estimates extracted from the resulting video (blue superimposed cuboid representing estimated 3D rotation) and crops the movement, setting precise start and stop points in time; (c) Approximate matches in target videos are found, ranked on similarity, and displayed, and the PT can watch them back, consider context, and flag them with a range of different colour codes based on their relevance (here red and green flags have been used); (d) The PT can view graphical search summaries showing the distribution of flagged events over the duration of the programme of therapy, and click on them to access the original movements in the target videos, along with date and time information (by clicking individual histogram elements in each interval).

2.2.3. Ranking Results

Results are scored and ranked based on their discrete Fréchet distance (DFD) (Eiter & Mannila, 1994) from the PT's original query data, S. The DFD is given by the minimum distance required to connect a point moving through the consecutive samples in the query data with a point moving through the consecutive samples in the result, but where the rate of movement of each point need not be uniform. Typically, the distance measure is a Euclidean distance between points in an N-D space, but here we use Equation (1) to compare distances between quaternions. The result with the lowest DFD score (closest match to S) receives the top ranking, and so on.

2.2.4. Generating Subsequent Results

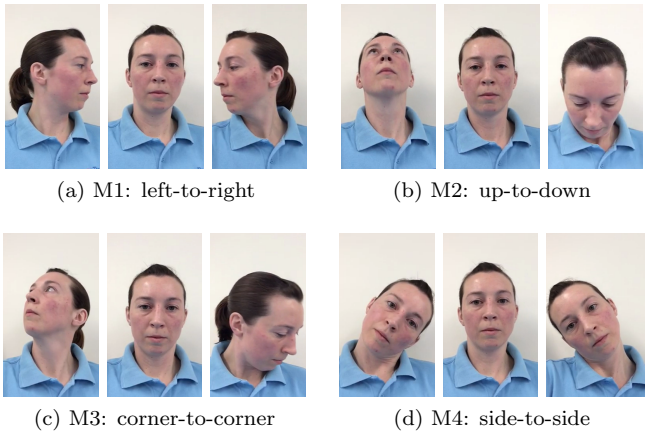
Extra 'pages' of results can be generated by increasing δ to include more clusters in the sets C_{start} and C_{stop}, and repeating the result generation steps above, subject to the additional condition that any result which overlaps with one already seen on a previous page, is deleted.

2.3. Search System

A browser-based search system was designed to allow PTs to extract head poses from query videos of themselves and from the target video using the OpenFace face tracker (Baltrušaitis et al., 2016, 2013), and then compare the two sets of results using the methods defined in the previous section. The system resulted from a participatory design process (Schuler & Namioka, 1993), where PTs tested and gave feedback on an initial prototype to which investigators then responded. An overview of the final system is given in Fig. 4.

2.3.1. Recording a Query Video

PTs can access a 'record' page to record new query videos via a webcam. This is done directly through the browser window, using the `MediaRecorder` API (Mozilla, 2018), and the search system is intended for use on a standard laptop or tablet at the clinic (Fig. 4a). PTs can use the 'record' page to review, re-record and/or save as many different query videos as they need.



(a) M1: left-to-right (b) M2: up-to-down

(c) M3: corner-to-corner (d) M4: side-to-side

Figure 5: Examples of each of the four head movements M1-M4 (a-d) searched for by the PTs.

2.3.2. Extracting Pose Estimates

Once query videos have been recorded, PTs access an 'update' page, which makes asynchronous requests to a local Node.js server in order to process the videos with the OpenFace face tracker (Baltrušaitis et al., 2016, 2013) and cluster the resulting head pose estimates. The 'update' page also handles the processing of any new target videos that are copied onto the computer, extracting and clustering pose estimates in the same way. The 'update' page requires no user interaction and is designed to be left permanently running in the clinic. Once processing of a new query video is complete, a 'crop' page enables PTs to manually step though the individual frames of their video with the OpenFace pose estimates superimposed, setting new start and stop times (Fig. 4b). This process also allows PTs to review the accuracy of the pose estimation, and discard a query video that has produced errors.

2.3.3. Searching Target Videos

PTs' query videos are saved permanently and can be used to search a collection of target videos from any one individual with a single button press. Each search result is represented by a static video abstract comprised of 4 images, evenly spaced in time across the duration of the corresponding movement. Extra pages of results can be generated by clicking a 'next' button, and PTs are free to generate and navigate between as many pages of results as they wish. Clicking on a static abstract causes a video playback modal to appear through which the video of the corresponding movement can be played back and, if desired, flagged with colour-coded flags. No fixed meanings are attached to flag colours, with PTs free to use them as they wish. Upon applying a flag the original static abstract is highlighted in the same colour (Fig. 4c). Watched but unflagged results are set translucent to help the PT keep track of their progress.

Search results are stored permanently and a PT can access and resume them at any future time. If new target videos have been added then, upon resuming an existing search, any new results automatically appear within the overall list of results (ranked appropriately). In this way PTs can retain searches for long periods of time, returning to them to add flags as more target video data becomes available.

2.3.4. Graphical Overviews

PTs can access a graphical single-page summary for each of their searches, where target video timestamps and durations are used to plot the total number of minutes recorded each day, and a histogram of the total number of flagged events within each day (Fig. 4d). Clicking individual events in the histogram intervals causes the original movement from the target video to be played back, along with its date and time.

3. Experiments

3.1. Overview

Clinicians took cameras into their own homes to generate a structured target video dataset containing relevant head movement events. PTs subsequently searched their colleagues' target videos in order that the search system's performance could be evaluated.

3.2. Participants

Three PTs (PT1-PT3, all female, based at The Movement Centre, Oswestry, UK) participated in the study. The PTs specialised in training movement control, including head control, in children with neuromotor disability. Their clinical experience in this field was between 3 and 14 years.

3.3. Target Video Datasets

The PTs each used a 2-camera Netgear Arlo (Netgear, 2018) to film themselves during three separate recording sessions, each lasting no less than 30 minutes. The first session was recorded at the clinic, and the following two sessions were recorded in the PT's own home. At times of their own choosing during each session, PTs performed controlled left-to-right head rotations (movement M1, Fig. 5a), as an example of a clinically relevant patient movement. One camera was placed directly in front of them to give a view of their face allowing for head pose estimation, and one camera placed to their left or right to give a side-view of the full upper body for additional context. The total number of relevant movements performed was known *only to that PT* and passed to one of the investigators for use in evaluating the subsequent search tasks. In total, the three PTs recorded just over 3.5 hours of data.

One investigator then used the Arlo system at home to record: (i) a '5-day' dataset, identical in design to the PT dataset, but featuring five longer sessions on each of 5 consecutive days (total of 4.2 hours), and including examples of all of the movements M1-M4 from Fig. 5; (ii) an 'Empty' data set (total of 55 minutes) containing no relevant head movements. The nature of both these datasets was completely unknown to the PTs.

3.4. Parameter Settings

Pose estimates from all of the target video datasets were extracted and clustered using the 'update' page, before PTs attempted any search tasks. $N = 5000$ cluster centres were used to generate the runs $\{\mathcal{R}_i\} \forall i \in C$. A value for δ was estimated from recordings of the participants during comfortable still sitting, by retaining the maximum observed change in head rotation, $\delta = 0.1162$. This value was scaled in integer multiples to generate subsequent pages of search results $(2\delta, 3\delta, ...)$.

3.5. Search Tasks

The three PTs were asked to record two query videos each and complete the following search tasks: (i) finding every example of M1 for each of their colleagues; (ii) finding every example of M1 in the '5-day' dataset; (iii) finding every example of one other movement (M2, M3, or M4) in the '5-day' dataset; (iv) finding every example of M1 in the 'Empty' dataset (though in fact there were none).

3.6. Analysis

For each search, the following metrics were computed:

Events found: The number of relevant events flagged by the searching PT, versus the true number of relevant events contained in the target videos.

Search times: The time taken from a PT clicking onto the first page of results until they reported having finished their search.

Time savings: The difference between the *search time* and the time that would be required to watch the target videos back in full, expressed as a percentage of the latter.

Table 1: Search results: PTs each searched all of their colleagues' target videos for the movement M1.

	Events Found			Search Time (mm:ss)			Time Saving (%)		
Search task	PT1	PT2	PT3	PT1	PT2	PT3	PT1	PT2	PT3
PT1(M1)	-	26/27	6/6	-	18:18	6:26	-	83.6	89.8
PT2(M1)	20/21	-	6/6	17:15	-	7:50	82.0	-	87.6
PT3(M1)	19/21	24/27	-	8:12	11:15	-	91.4	89.9	-

Table 2: Search results for the '5-day' target video dataset. This dataset contains over 4.2 hours of video in total.

Search task	Events Found	Search Time (mm:ss)	Time Saving (%)
PT1(M1)	34/35	13:26	94.7
PT1(M2)	2/2	3:30	98.6
PT2(M1)	34/35	15:13	94.0
PT2(M3)	4/4	4:22	98.3
PT3(M1)	35/35	10:39	95.8
PT3(M4)	3/3	2:12	99.1

4. Results

Table 1 shows the number of events found, search time and estimated time savings for each of the PTs searching their colleagues' target videos. We use the notation PTN(M#) to refer to the Nth PT's search for the movement M#. Table 2 and Table 3 show the same three measures for the PTs' searches of the '5-day' and 'Empty' datasets, respectively.

Fig. 6 shows every PT's search time for every target video dataset, versus the true number of relevant events contained in the corresponding target videos.

From a clinical perspective, in total across all their searches, the PTs found 229/238 events (or 96.2%). In almost every case, relevant events that were missed were due to pose estimation failures on the target videos. OpenFace pose estimation can fail if facial features are substantially occluded and most failures (and missed events) came from self-occlusions due to rotation of the head close to or past 90° relative to the camera. Pose estimation failures on query videos were not an issue because PTs saw pose estimation results when cropping their query videos and could identify problems and re-record if necessary. (Though this was not necessary in the evaluation here.)

Search times did not appear to be related to the durations of the target video datasets (Table 2) but increased with the true number of relevant events in the target videos (Fig. 6). This result is encouraging as in a real clinical deployment, PTs would likely be searching very many/long target videos for relatively few relevant events. The time savings for all individual searches were above 80% with most above 90%, and many of the savings on the longer '5-day' dataset above 95%. For example, PT3 found all 35/35 left-to-right (M1) events in 10 min 39 sec, a time saving of 95.8%, and all 3/3 side-to-side (M4) events in just 2 min 12 sec, a time saving of 99.1%. These time savings are probably a conservative estimate, since conventional searching of a video for specific events involves repeated replaying of sections of that video.

Table 3: Search results for the 'Empty' target video dataset (55 minutes of video).

Search task	Events Found	Search Time (mm:ss)	Time Saving (%)
PT1(M1)	0/0	1:54	96.6
PT2(M1)	0/0	1:37	97.1
PT3(M1)	0/0	2:33	95.4

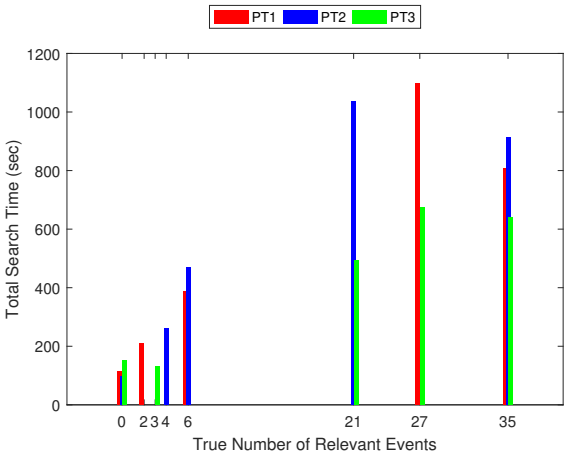


Figure 6: Search times versus the true number of relevant events contained in the target videos, for every search by every PT.

5. Discussion

Although there is undoubtedly further work to be undertaken, we believe we have created a first generation video-based retrospective system to address the objectives we defined of:

i) enabling any patient to share information about progress at home, simply by sharing video (with assistance from carers, if needed) and

ii) building intelligent systems to support Physical Therapists (PTs) in reviewing the resulting video data and extracting the detail relating to the body movements of interest for a given patient: i.e., video-based retrospective systems.

This proposed system potentially enables PTs to achieve improved healthcare outcomes for their patients.

5.1. Discussion of the System

In contrast to other retrospective systems (e.g., SenseCap (Huang et al., 2014) and ArmSleeve (Ploderer et al., 2016b)), our proposed system is unique in allowing PTs to i) look for specific user-defined fine-grained movements, ii) providing them with video for context, and iii) providing summaries of their previous searches. In comparison with other fine-grained action recognition approaches (e.g., Chéron et al. (2015)), our approach has the advantage that it does not require large amounts of training data for new action classes, needing only a single example from the PT. In addition, because pose-based features are easy to interpret (versus appearance- or motion-based features) PTs can review their queries to ensure they accurately capture the information they require for clinical judgement (Section 2.3.2). The algorithm for comparing queries with longer videos reliably produces high recall results but with relatively low precision, consistent with existing fully automatic content-based video retrieval approaches. However, PTs can use the interactive search interface to significantly increase precision without sacrificing recall, and in a substantially shorter time than would be required to review the footage in full (with reductions of as much as 99.1%). In common with other interactive search methods, PTs can also use the system to browse movements where their clinical interests are less well defined. As an example, rather than stopping the search as soon as movements begin to depart significantly from their original search query, they can gain valuable clinical information by continuing their search to see what form these departures take and their relative rates of occurrence (e.g., where a patient is struggling with a particular exercise or a commonly performed movement). In contrast to other content-based video retrieval methods, the system is able to retrieve fine-grained human movements, important to the PT needs, where contemporary concept-based systems (e.g., Ueki et al. (2017)) support only coarse actions such as 'stand up', and 'walk'. The system is also the first content-based video retrieval

system to support query-by-movement, allowing for the natural and unambiguous definition of new queries for fine-grained movements (e.g., versus typing text-based descriptions).

The potential benefits of adopting the system include the provision of feedback to the patient or carer on a rapid basis, using a conventional telemedicine approach if needed. It also opens up a series of associated benefits, including better-informed decision-making when updating the plan of care which can be done prior to the next face-to-face session and the ability to schedule face-to-face sessions more responsively and more efficiently (e.g., delaying or expediting them based on patient progress). All these have positive implications for eventual patient outcomes, PT time, and cost, over the course of a programme of therapy.

5.1.1. Limitations of the System

Our approach is not as robust to variations in video footage as contemporary methods for coarse action recognition (e.g., Simonyan & Zisserman (2014)) and content-based video retrieval (e.g., Ueki et al. (2017)). We assume only partial occlusions and good lighting for pose estimation to succeed, and static cameras facing towards the patient/clinician in order to achieve consistent measurements of relative head rotation that are suitable for comparison. Pose estimation itself is relatively resource intensive, requiring reasonable processing power and storage space to be available at the clinic, with associated cost implications (OpenFace runs in approximately real-time, requiring 1 minute of processing time to extract pose estimates from 1 minute of video). However, it is this shifting of the processing burden for pose estimation into the clinic that ensures patients are free to share video in any way they wish, and without the need for bespoke hardware/software in their home (as necessary with all other proactive and retrospective systems). Using the system to achieve good recall rates takes time (e.g., the longest search conducted by any PT in this study was just over 18 minutes), again with resource implications for the clinic. Cost implications for the clinic are discussed further in Section 5.3. Finally, if pose estimation fails – either during a video recorded at home, or a query recorded at the clinic – then search will fail. However, the performance and rate of improvement in modern monocular head pose estimation methods is impressive and our proposed system is independent of the particular method used, meaning that new methods can be easily substituted in as they become available (e.g., substituting OpenFace for OpenFace 2.0 (Baltrusaitis et al., 2018)).

The algorithm underlying the proposed system is a relatively simple strategy for the cluster-based identification and DFD-based ranking of similar sequences of head poses. We used naive clustering of quaternions, comparing every head pose against every cluster centre using Equation (1) to find nearest neighbours. For a large amount of data like the '5-day' dataset this took approximately 20 minutes. This time could be reduced through space partitioning methods such as k-d trees or R-trees (Guttman, 1984). The DFD was used to compare those sequences identified through clustering, with a PT's query, and rank them. However, this comparison does not take account of differences in the speed of movements and this would be a valuable addition. More sophisticated algorithms for sequence comparison that can be used to account for differences in time (e.g., Dynamic Time Warping) should be considered in future work (see also Section 5.3). Finally, versus other interactive methods for content-based video retrieval, the proposed system does not yet allow for any form of *relevance feedback* (Rossetto et al., 2015; Lu et al., 2017), where the user's interactions with the system are continually fed back into the underlying search mechanism (e.g., by updating an original query to incorporate results marked as relevant); specific suggestions for future work in this area are made in Section 5.3.

5.2. Discussion of the Study

We believe the target video datasets, with their ground truth completely unknown to the PTs searching them, allow a meaningful and positive picture of clinicians' trust in search results to emerge. The relationship in Fig. 6 continued to hold true for the 'Empty' dataset, which contained no relevant events, with PTs confident to terminate their search. This finding, together with the significant time savings and assurance of finding relevant events gives confidence to pursuing this search approach as a clinically appropriate and useful system.

We took a clinician-led approach to the early stages of development: this investment by the PTs helped to shape the software and enabled them to have first-hand experience of the hardware. We believe that this

was vital to the success of this study and will be critical to proposed future work with patients (see Section 5.3).

This clinician-led participatory design process not only helps ensure that software does what is clinically required, but helps to build a sense of ownership that could avoid previously reported perceptions of 'technology push' (European Commission, 2014). It also ensures clinicians have the depth of knowledge necessary to provide high quality guidance to their patients (Hale & Kvedar, 2014) such as ensuring consent is properly informed (McCall & Baillie, 2017) around complex issues such as privacy and data security (Hall & McGraw, 2014).

5.2.1. Limitations of the Study

This was a relatively small study, involving three PTs and focusing on four different head movements. Although there are advantages in the long term to starting with a clinician-led study (as above), the patient's experience of the system has not been measured in this present study and remains an important topic for future investigation. This would include the security concerns that surround sharing video data recorded at home (Climent-Pérez et al., 2019). Once these issues have been addressed, studies can investigate the impact on patient progress or efficiency savings during a real programme of physical therapy.

5.3. Implications and Future Work

The following sections discuss the implications of the study and propose future work to improve the current state of knowledge in the field, both from a clinical and technical perspective.

5.3.1. Clinical Perspective

Our proposed system has the potential for immediate use in the retrospective review of patient movement at home, including movements by patients who are not readily able to engage with proactive systems requiring interaction, or who are not comfortable wearing invasive sensors. The query-by-movement approach has considerable possible impact since PTs have freedom to investigate whatever movements they need to, and without the need for any technical assistance. This opens this approach to application in numerous different forms of therapy, searching for unique movements relevant to a particular patient's disability or progress. To date, the customisation of movement specifications to individual patients has only been demonstrated in proactive systems, and only possible with technical assistance (Alankus et al., 2010).

Our proposed next stage will be to support a PT-led deployment of Arlo cameras to child patients, and using the system to review shared videos. Although the system is not tied to any particular method of recording video, cameras systems such as the Arlo are completely wireless and can be attached to a patient's standing frame or other device used specifically for head control training. They can also be transferred to other equipment used by the patient, such as a wheelchair, or used to record video of daily functional activities.

This study used a clinicians first approach (Hochstenbach-Waelen & Seelen, 2012) with the result that the participating PTs are well placed to advise patients and their families on the potential benefits of the system and the realities of taking a Netgear Arlo camera home. However, to promote wider consideration/adoption of the system, the following resources would be valuable: i) a repository of information on other camera hardware that can be used to share video from home; ii) guidance to patients on recording good quality footage (i.e., where subsequent pose estimation at the clinic is likely to succeed); iii) packaging of new and alternative pose estimation software for easy download and use with the system by clinics (e.g., new versions of OpenFace, or future alternatives); iv) performance metrics from larger clinician-led studies investigating a wider range of movements relevant in different therapies.

The practical implications of adopting the system for a clinic include: the cost of camera hardware for patients; the cost of computing and storage hardware for the clinic; the cost of staff time spent on reviewing videos. Each of these implications deserves proper consideration in future work against the existing costs and efficacy of a programme of therapy where progress sharing from home is not possible. No such comparison has been published for any system for the retrospective clinical review of human movement at home, and is important outstanding work.

5.3.2. Technical Perspective

Larger clinician-led studies would offer an opportunity for the further investigation of a number of technical implications. Including, first, the effect of the sequence comparison algorithm on rankings and search times. Alternatives to the DFD such as Geometric Edit Distance and Dynamic Time Warping could be explored. The latter could also allow for the consideration of temporal differences between movements using recovered warping paths, which may be valuable where the speed of movement is an important clinical aspect. Second, other pose estimation approaches (e.g., Zhang et al. (2014); Asthana et al. (2014)) could be evaluated both in terms of search performance metrics but also the associated computational overhead at the clinic. The latter should consider the use of space partitioning methods such as k-d trees and R-trees (Guttman, 1984) during the subsequent clustering of pose estimates. Third, large clinician-led studies would allow exploration of the effect of relevance feedback strategies on search performance metrics. Specifically, we would suggest allowing PTs to dynamically substitute relevant results for their initial query, or add them to a *set* of queries and compute minimum DFD scores across the set.

Other studies of retrospective systems (Huang et al., 2014; Ploderer et al., 2016b,a) have proposed sharing visual summaries with patients and carers to provide motivation and feedback on progress. The methods proposed in this study ensure that video data only ever flows from patient to PT, ensuring that patients and clinics can benefit from the secure infrastructure offered by camera manufacturers. (E.g., Netgear use AES 128-bit encryption of video, Transport Layer Security (TLS) for upload, sharing by invitation only, enforcement of strong password requirements, and the use of a secure HTTPS connection for download.) However, future work could look at methods for the secure sharing of search summaries from the clinic to the home (e.g., using secure data transmission protocols via a web browser) with large clinician-led studies evaluating the resulting benefits.

Finally, although we have used the example of head pose, applications of the methods described are possible wherever there exists an appropriate image-based pose estimation algorithm for the body parts of interest, e.g. fingers, face, arm, etc. The methods described in Section 2 are applicable to other kinds of pose-based features by substituting Equation (1) for a suitable distance metric for use in the initial clustering and the computation of DFD scores. Euclidean distance could be used, for example, to compare joint locations or facial landmarks. Applications to other pose-based features relevant to other types of physical therapy would be a valuable topic for future investigation.

Applying interactive content-based video retrieval methods to finding fine-grained human actions is a novel contribution of the work. Answering these outstanding technical questions would allow an initial performance benchmark to be established, and encourage research and contributions on the topic, similarly to those benchmarks established through events like the Text Retrieval Conference's Video Retrieval Evaluation (Awad et al., 2018) and the Video Browser Showdown Lokoč et al. (2019).

6. Conclusions

In this study, we have used a clinician-led approach to develop an interactive content-based video retrieval system based on contemporary image-based pose estimation techniques, thus enabling any patient to share video of their movements at home. This work contributes the first retrospective progress monitoring system to preserve context through video, the first content-based video retrieval method to support query-by-movement and the first content-based video retrieval method to support fine-grained human actions. An application to head pose has demonstrated excellent performance in terms of recall and search time, while the iterative, user-centred design process has ensured the associated interface meets the needs of Physical Therapists (PTs). Our results provide support to sharing videos as a promising way for patients to provide their PTs with information about progress made at home and information that may otherwise be unavailable due to patient constraints such as age and ability to comply. Video sharing is non-invasive, preserves the context of movements for the PT to see, and allows for sharing information about movement during daily life in addition to movement during recommended activity. The disadvantages are the initial costs of camera hardware for the home, and processing and storage hardware for the clinic. We propose future research, including large clinician-led studies, to explore these areas further and to refine some of the current technical

aspects, leading to greater benefits in terms of patient outcomes, PT time, and reduction in cost, over the course of a programme of therapy.

Acknowledgment

We are grateful to staff at The Movement Centre who contributed test data and/or provided feedback on prototypes, in particular Pauline Holbrook and Lynne Ford and also to Richard Major.

References

Aggarwal, D., Ploderer, B., Vetere, F., Bradford, M., & Hoang, T. (2016). Doctor, can you see my squats?: Understanding bodily communication in video consultations for physiotherapy. In *Proceedings of the ACM Conference on Designing Interactive Systems* (pp. 1197–1208). ACM.

Alankus, G., & Kelleher, C. (2015). Reducing compensatory motions in motion-based video games for stroke rehabilitation. *Human–Computer Interaction*, *30*, 232–262.

Alankus, G., Lazar, A., May, M., & Kelleher, C. (2010). Towards customizable games for stroke rehabilitation. In *SIGCHI Conference on Human Factors in Computing Systems* (pp. 2113–2122).

Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2014). Incremental face alignment in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1859–1866).

Awad, G., Butt, A., Curtis, K., Lee, Y., Fiscus, J., Godil, A., Joy, D., Delgado, A., Smeaton, A., Graham, Y. et al. (2018). Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search.

Awad, G., Butt, A., Fiscus, J., Joy, D., Delgado, A., Mcclinton, W., Michel, M., Smeaton, A., Graham, Y., Kraaij, W. et al. (2017). Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning, and hyperlinking.

Bailer, W., Weiss, W., & Wechtitsch, S. (2016). Selecting user generated content for use in media productions. In *International Conference on Multimedia Modeling* (pp. 388–393). Springer.

Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 354–361). IEEE.

Baltrušaitis, T., Robinson, P., Morency, L.-P. et al. (2016). OpenFace: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision* (pp. 1–10). IEEE.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 59–66). IEEE.

Barthel, K. U., Hezel, N., & Mackowiak, R. (2015). Graph-based browsing for large video collections. In *International Conference on Multimedia Modeling* (pp. 237–242). Springer.

Chéron, G., Laptev, I., & Schmid, C. (2015). P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 3218–3226).

Choutas, V., Weinzaepfel, P., Revaud, J., & Schmid, C. (2018). Potion: Pose motion representation for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7024–7033).

Climent-Pérez, P., Spinsante, S., Michailidis, A., & Florez-Revuelta, F. (2019). A review on video-based active and assisted living technologies for automated lifelogging. *Expert Systems with Applications*, (p. 112847).

Eiter, T., & Mannila, H. (1994). *Computing discrete Fréchet distance*. Technical Report CD-TR 94/64 Information Systems Department, Technical University of Vienna.

European Commission (2014). Green paper consultation on mobile health. https://ec.europa.eu/digital-single-market/en/news/green-paper-mobile-health-mhealth. Accessed: 2019-25-06.

Fanelli, G., Gall, J., & Van Gool, L. (2012). Real time 3D head pose estimation: Recent achievements and future challenges. In *International Symposium on Communications Control and Signal Processing* (pp. 1–4).

Geetha, P., & Narayanan, V. (2008). A survey of content-based video retrieval, .

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27–48.

Guttman, A. (1984). *R-trees: A dynamic index structure for spatial searching* volume 14. ACM.

Hale, T., & Kvedar, J. (2014). Privacy and security concerns in telehealth. *The virtual mentor*, *16*, 981.

Hall, J. L., & McGraw, D. (2014). For telehealth to succeed, privacy and security risks must be identified and addressed. *Health Affairs*, *33*, 216–221.

Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. *Image and vision computing*, *60*, 4–21.

Hochstenbach-Waelen, A., & Seelen, H. A. (2012). Embracing change: practical and theoretical considerations for successful implementation of technology assisting upper limb training in stroke. *Journal of neuroengineering and rehabilitation*, *9*, 52.

Hu, W., Xie, D., Fu, Z., Zeng, W., & Maybank, S. (2007). Semantic-based surveillance video retrieval. *IEEE Transactions on image processing*, *16*, 1168–1181.

Hu, W., Xie, N., Li, L., Zeng, X., & Maybank, S. (2011). A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *41*, 797–819.

Huang, K., Sparto, P. J., Kiesler, S., Smailagic, A., Mankoff, J., & Siewiorek, D. (2014). A technology probe of wearable in-home computer-assisted physical therapy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2541–2550). ACM.

Hürst, W., van de Werken, R., & Hoet, M. (2015). A storyboard-based interface for mobile video browsing. In *International Conference on Multimedia Modeling* (pp. 261–265). Springer.

Huynh, D. Q. (2009). Metrics for 3D rotations: Comparison and analysis. *J. Math. Imaging Vis.*, *35*, 155–164.

Kimel, J. C. (2005). Thera-network: A wearable computing network to motivate exercise in patients undergoing physical therapy. In *25th IEEE International Conference on Distributed Computing Systems Workshops* (pp. 491–495). IEEE.

Kläser, A., Marszałek, M., Schmid, C., & Zisserman, A. (2010). Human focused action localization in video. In *European Conference on Computer Vision* (pp. 219–233). Springer.

Lokoč, J., Bailer, W., Schoeffmann, K., Muenzer, B., & Awad, G. (2018). On influential trends in interactive video retrieval: video browser showdown 2015–2017. *IEEE Transactions on Multimedia*, *20*, 3361–3376.

Lokoč, J., Kovalčík, G., Münzer, B., Schöffmann, K., Bailer, W., Gasser, R., Vrochidis, S., Nguyen, P. A., Rujikietgumjorn, S., & Barthel, K. U. (2019). Interactive search or sequential browsing? a detailed analysis of the video browser showdown 2018. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *15*, 29.

Lu, Y.-J., Nguyen, P. A., Zhang, H., & Ngo, C.-W. (2017). Concept-based interactive search system. In *International Conference on Multimedia Modeling* (pp. 463–468). Springer.

Luvizon, D. C., Picard, D., & Tabia, H. (2018). 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5137–5146).

McCall, R., & Baillie, L. (2017). Ethics, privacy and trust in serious games. *Handbook of Digital Games and Entertainment Technologies*, (pp. 611–640).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Moreno-Schneider, J., Martínez, P., & Martínez-Fernández, J. L. (2017). Combining heterogeneous sources in an interactive multimedia content retrieval model. *Expert Systems with Applications*, *69*, 201–213.

Mortazavi, B. J., Pourhomayoun, M., Lee, S. I., Nyamathi, S., Wu, B., & Sarrafzadeh, M. (2016). User-optimized activity recognition for exergaming. *Pervasive and Mobile Computing*, *26*, 3–16.

Moumtzidou, A., Mironidis, T., Markatopoulou, F., Andreadis, S., Gialampoukidis, I., Galanopoulos, D., Ioannidou, A., Vrochidis, S., Mezaris, V., Kompatsiaris, I. et al. (2017). Verge in vbs 2017. In *International Conference on Multimedia Modeling* (pp. 486–492). Springer.

Mozilla (2018). MediaRecorder API. https://developer.mozilla.org/en-US/docs/Web/API/MediaRecorder. Accessed: 2018-07-09.

Murphy-Chutorian, E., & Trivedi, M. M. (2009). Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*, 607–626.

Negin, F., Rodriguez, P., Koperski, M., Kerboua, A., Gonzàlez, J., Bourgeois, J., Chapoulie, E., Robert, P., & Bremond, F. (2018). Praxis: Towards automatic cognitive assessment using gesture recognition. *Expert systems with applications*, *106*, 21–35.

Netgear (2018). Netgear Arlo. https://www.arlo.com/. Accessed: 09-07-2018.

Ni, B., Paramathayalan, V. R., & Moulin, P. (2014). Multiple granularity analysis for fine-grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 756–763).

Ploderer, B., Fong, J., Klaic, M., Nair, S., Vetere, F., Lizama, L. E. C., & Galea, M. P. (2016a). How therapists use visualizations of upper limb movement information from stroke patients: a qualitative study with simulated information. *JMIR rehabilitation and assistive technologies*, *3*, e9.

Ploderer, B., Fong, J., Withana, A., Klaic, M., Nair, S., Crocher, V., Vetere, F., & Nanayakkara, S. (2016b). Armsleeve: a patient monitoring system to support occupational therapists in stroke rehabilitation. In *Proceedings of the 2016 Conference on Designing Interactive Systems* (pp. 700–711). ACM.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, *28*, 976–990.

Rohrbach, M., Amin, S., Andriluka, M., & Schiele, B. (2012). A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1194–1201). IEEE.

Rossetto, L., Giangreco, I., Schuldt, H., Dupont, S., Seddati, O., Sezgin, M., & Sahillioğlu, Y. (2015). Imotion—a content-based video retrieval engine. In *International Conference on Multimedia Modeling* (pp. 255–260). Springer.

Saini, S., Rambli, D. R. A., Sulaiman, S., Zakaria, M. N., & Shukri, S. R. M. (2012). A low-cost game framework for a home-based stroke rehabilitation system. In *Computer & Information Science (ICCIS), 2012 International Conference on* (pp. 55–60). IEEE volume 1.

Schoeffmann, K., Hopfgartner, F., Marques, O., Boeszoermenyi, L., & Jose, J. M. (2010). Video browsing interfaces and applications: a review. *SPIE Reviews*, *1*, 018004.

Schoeffmann, K., Hudelist, M. A., & Huber, J. (2015). Video interaction tools: a survey of recent work. *ACM Computing Surveys (CSUR)*, *48*, 14.

Schuler, D., & Namioka, A. (1993). *Participatory design: Principles and practices*. CRC Press.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems* (pp. 568–576).

Sivic, J., & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision* (pp. 1470–1477). IEEE.

Snoek, C. G., Worring, M., Koelma, D. C., & Smeulders, A. W. (2007). A learned lexicon-driven paradigm for interactive video retrieval. *IEEE Transactions on Multimedia*, *9*, 280–292.

Soomro, K., Zamir, A. R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, .

Tang, R., Yang, X.-D., Bateman, S., Jorge, J., & Tang, A. (2015). Physio home: Exploring visual guidance and feedback techniques for physiotherapy exercises. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 4123–4132).

Ueki, K., Hirakawa, K., Kikuchi, K., Ogawa, T., & Kobayashi, T. (2017). Waseda meisei at trecvid 2017: Ad-hoc video search. In *Proc. of TRECVID 2017*.

Ueki, K., Kikuchi, K., Saito, S., & Kobayashi, T. (2016). Waseda at trecvid 2016: ad-hoc video search. In *TRECVID 2016 Workshop, Gaithersburg, MD, USA* (pp. 110–111). volume 24.

Uzor, S., & Baillie, L. (2013). Exploring & designing tools to enhance falls rehabilitation in the home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1233–1242). ACM.

Yang, L., Cai, Y., Hanjalic, A., Hua, X.-S., & Li, S. (2013). Searching for images by video. *International Journal of Multimedia Information Retrieval*, *2*, 213–225.

Yao, A., Gall, J., Fanelli, G., & Van Gool, L. J. (2011). Does human action recognition benefit from pose estimation?. In *BMVC* (p. 6). volume 3.

Yao, A., Gall, J., & Van Gool, L. (2010). A hough transform-based voting framework for action recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 2061–2068). IEEE.

Yershova, A., Jain, S., Lavalle, S. M., & Mitchell, J. C. (2009). Generating uniform incremental grids on SO(3) using the Hopf fibration. *International Journal of Robotics Research*, *29*, 801–812.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4694–4702).

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2014). Facial landmark detection by deep multi-task learning. In *European conference on computer vision* (pp. 94–108). Springer.